Summarizing Video Collections using Semantic Graph

Philippe Ercolessi*, Christine Sénac*, Hervé Bredin†, Philippe Joly*
*IRIT, Universit Paul Sabatier, Toulouse, France
†LIMSI, Orsay, France

I. Introduction

As the amount of available digital video content is increasing exponentially, novel ways of storing, accessing and retrieving it are being developped, such as indexing, segmentation or abstraction techniques. Video abstraction can be useful in many ways - from automatic home movie editing to easier (and faster) exploration of a video collection. Video summaries can be a set of carefully selected key frames or the concatenation of video skims [1].

Rather than focusing on a single video, we aim at summarizing several videos from a single collection. A video collection can be seen as a homogeneous set of videos, with the same type of structure, style, duration, etc. Given a video collection, the objective is to generate a short video that is a representative of this collection. The representativeness of a video depends on the application (faster browsing or narrative summary, for instance). It is not reasonable to process each element of a collection independently to produce the final collection summary. As a matter of fact, such an approach would not take redundancy beween videos into account and there is no way it can uncover the commonalities between episodes. Therefore, our approach considers a collection as a whole and rely on two steps : video structuring and structure comparison.

Considering this work, we focus on collections made of american TV series such as *Ally McBeal* and *Malcolm in the Middle*.

II. Video structuring

A. Multimodal segmentation

Videos are segmented based on various visual and audio descriptors. For instance, videos can be devided into *Logical Story Units* (LSU) [1] based on color histograms, speaker diarization [3] and automatic speach recognition [2].

B. Structuring

Considering that we are focusing on strongly narrative videos with a highly intricate plot, we define the structure of a video by a set of stories and substories tangled together.

Thus, the structuring step consists in grouping together similar LSU which are parts of a same story. This is made by clustering using color, speakers presence and speach recognition.

III. Structure comparison

The comparison step consists in finding similarities between the structure of each episode of the collection. To compare these structures, the clustering step is applied on LSUs of the whole collection. Semantically similar segments are gathered in the same cluster and temporal relationships between them lead to the definition of one graph per video, with one cluster per node of the graph.

We use these graphs to detect what could be considered as a common characteristic of the whole collection, or to the contrary, what is video-specific (see figure 1). This automatic process has to run without any a priori knowledge on the nature of the collection. Therefore, methods as generic as possible are implemented.

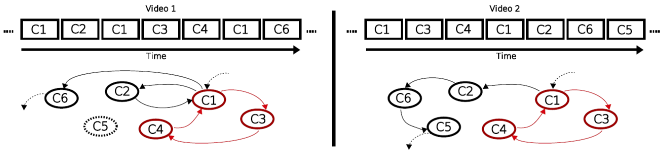

Fig. 1. Building and comparing two semantic graphs. After the initial segmentation step, each video segment is associated to the closest cluster C_i. Semantic graphs are then build by simply linking temporally consecutive segments. Red clusters and links constitute a sub-graph that is common to both videos. Cluster C_5 is absent in the first video.

IV. Collection abstraction

Once common characteristics or specific parts of a collection are clearly determined, it is then possible to generate a video summary which will focus on reccuring parts of the collection, or on specific parts of each video (in the case when one summary per video is needed).

References

[1] Sergio Benini, Pierangelo Migliorati, and Ricardo Leonardi, "Statistical Skimming of Feature Films." in *International Journal of Digital Multimedia Broadcasting*, vol. 2020, Hindawi Publishing Corporation, 2009.

[2] Elie El Khoury, Christine Sénac, and Philippe Joly. "Face-and-clothing based People Cluster- ing in Video Content." In *MIR '10: Proceedings of the International Conference on Multimedia Information Retrieval*, pages 295-304, New York, NY, USA, 2010. ACM.

[3] Philippe Ercolessi, Hervé Bredin, Christine Sénac, and Philippe Joly, "Segmenting TV Series into Scenes Using Speaker Diarization." in *WIAMIS*, 2011.